Hybrid Active Learning via Deep Clustering for Video Action Detection

Aayush J Rana
aayushjr@knights.ucf.edu

Yogesh S Rawat
yogesh@crcv.ucf.edu

Center for Research in Computer Vision (CRCV)
University of Central Florida

Abstract

*In this work, we focus on reducing the annotation cost for video action detection which requires costly frame-wise dense annotations. We study a novel hybrid active learning (AL) strategy which performs efficient labeling using both **intra-sample** and **inter-sample** selection. The intra-sample selection leads to labeling of fewer frames in a video as opposed to inter-sample selection which operates at video level. This hybrid strategy reduces the annotation cost from two different aspects leading to significant labeling cost reduction. The proposed approach utilize Clustering-Aware Uncertainty Scoring (CLAUS), a novel label acquisition strategy which relies on both **informativeness** and **diversity** for sample selection. We also propose a novel Spatio-Temporal Weighted (STeW) loss formulation, which helps in model training under limited annotations. The proposed approach is evaluated on UCF-101-24 and J-HMDB-21 datasets demonstrating its effectiveness in significantly reducing the annotation cost where it consistently outperforms other baselines. Project details available at https://tinyurl.com/hybridclaus*

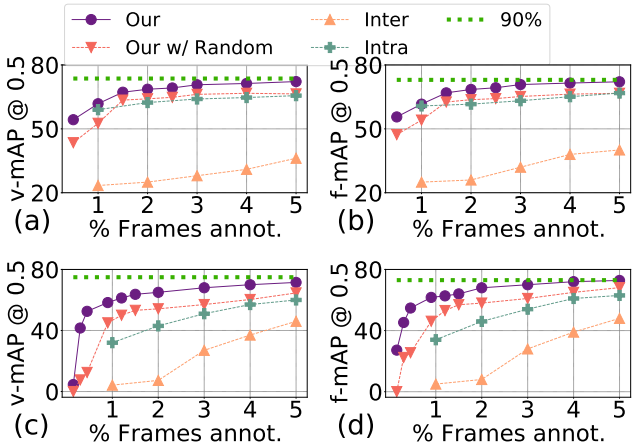

Figure 1. Comparison of the proposed CLAUS based AL method with random selection for video action detection. The plots show scores for (a-b) UCF-101-24 and (c-d) J-HMDB-21 for different annotation amount. The green line represents model performance with 90% annotations.

1. Introduction

Video action detection requires spatio-temporal annotations which include bounding-box or pixel-wise annotation on each frame of the video in addition to video level annotations [22, 27, 35, 50, 66]. Cost for such dense annotation is much higher compared to classification task where only video level annotations is sufficient [8, 18, 59, 60]. In this work, we study how this high annotation cost for spatio-temporal detection can be reduced with minimal performance trade-off.

The existing works on label efficient learning for videos are mainly focused on classification task [10, 23, 57]. Video action detection methods focus on weakly-supervised or semi-supervised methods to save annotation costs [12, 31, 39, 40, 64]. Weakly-supervised methods use partial annotations such as point-level [39], video-level [3, 12], and temporal annotations [9, 64]. Similarly, semi-supervised methods use unlabeled samples with the help of pseudo-labeling [40] and prediction consistency [31]. Such approaches have been effective for classification tasks, however spatio-temporal detection is more challenging under limited annotations with inferior performance compared to fully supervised methods. One of the main limitation of these methods is lack of selection criteria which can guide in labeling only informative samples. To overcome this limitation, we investigate the use of active learning for label efficient video action detection.

Traditional active learning (AL) approach typically focuses on classification task where selection is performed at sample level [19, 34, 61]. In video action detection, a frame-level spatio-temporal localization is required in addition to video level class prediction. Therefore, active learning strategy should also consider detection on every frame within a video apart from video-level decisions. We argue that frame-level informativeness is also crucial for spatio-temporal detection along with video-level informativeness. Motivated

by this, we explore a hybrid active learning strategy which performs both *intra-sample* and *inter-sample* selection. The intra-sample selection targets informative frames within a video and inter-sample selection aims at informative samples at video-level. This hybrid approach results in efficient labeling by significantly reducing the annotation costs.

Informativeness and diversity, both are important for sample selection in active learning [4]. The proposed approach utilize *Clustering-Aware Uncertainty Scoring (CLAUS)*, a novel clustering assisted AL strategy which considers both these aspects for sample selection. It relies on model uncertainty for informative sample selection and clustering for reducing redundancy. Clustering is jointly performed on feature space while model training where diversity is enforced based on cluster assignments. Moreover, the intra-sample selection will lead to limited annotations making model training difficult. To overcome this, we propose a novel training objective, *Spatio-Temporal Weighted (STeW)* loss, which relies on temporal continuity for pseudo labels and helps in learning under limited annotations.

We make the following contributions in this work: 1) ***novel hybrid AL strategy*** that selects frames and videos based on *informativeness* and *diversity*; 2) ***clustering*** based selection criteria that enables *diversity* in sample selection; 3) ***novel training objective*** for effective utilization of limited labels using *temporal continuity*. We evaluate the proposed approach on UCF-101-24 and JHMDB-21 and demonstrate that it outperforms other AL baselines and achieves comparable performance with model trained on 90% annotations at a fraction (5% vs 90%) of the annotation cost (Figure 1).

2. Related work

Video action detection: Recent advances in video action detection use video-specific modules for spatio-temporal feature understanding [15, 22, 27, 33, 42, 46, 66]. Earlier works extend image object detection techniques [36, 47, 49] to perform per frame detection and combine it with action classification, but they do not leverage temporal information while being computationally expensive. This led to 3D convolution based detection frameworks that combine tube/tubelets detection [15, 22, 66] with action classification [8, 18]. These methods use non-trivial multi-step region proposal for tube generation and classification. Recently, end-to-end method [11] with simpler training process have improved performance. All these methods rely on dense spatio-temporal annotation for performing well.

Limited label learning: Dense frame-wise spatio-temporal annotation is costly to obtain, therefore a natural step ahead was to use reduced annotations to train models for action detection task. Recent works on weakly and semi-supervised approach have shown comparable results in various tasks [2, 43, 52, 58] while reducing annotation cost drastically. The extension to action detection for weakly and semi-supervised approach [9, 12, 40, 63] has enabled using significantly reduced annotation compared to dense label based fully-supervised methods. These works only use video-level annotation [3, 12, 40, 67] or point-label or pseudo-label [9, 39, 63] but they do not have any criteria for selecting the limited samples and can spend annotation budget selecting redundant and non-informative samples.

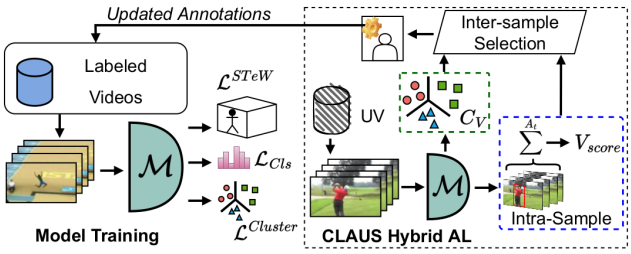

Figure 2. Overview of the proposed approach. Model training use videos with partial labels to learn action detection using the *STeW* loss and classification loss while also learning cluster assignment via cluster loss. The *CLAUS hybrid active learning* uses a trained model's output for intra sample selection and cluster assignment $\mathcal{C}_V$ for a video. Intra sample selection uses model score and selects top A_t frames of a video to get the video score (V_{score}). The V_{score} and $\mathcal{C}_V$ is used for inter sample selection and selected samples are sent to oracle for annotation. UV: Unlabeled videos.

Active learning: Iterative label assignment with AL to collect useful annotations is a widely used approach [19, 26, 34, 48, 55, 62, 65, 69], where a sample is selected based on its utility for the given task [13, 19, 34, 45, 69]. The utility function usually depends on uncertainty [14, 26, 38, 68], core-set selection [44, 54], clusters [4, 6], entropy [1, 20] or heuristics [30, 62]. These works mainly focus on image domain and lack formulation for temporal correlation existing in videos. There are some efforts focusing on detection [1] and instance segmentation [13], but they do not explore label sparsity and temporal aspect of videos. Moreover, these works are based on just informativeness property of samples without considering their diversity. We focus on spatio-temporal detection and explore a hybrid approach which considers both informativeness as well as diversity.

3. Methodology

Given a set of N videos $\mathcal{V} = \{v_1, v_2, ...v_N\}$ with $\mathcal{F}$ total frames, we intend to select a subset of videos $\mathcal{V}_s^T \subset \mathcal{V}$ with $\mathcal{F}_s^T$ frames and annotate $\mathcal{A}^{\mathcal{T}}\%$ of frames from the subset $\mathcal{V}_s^T$ based on the total budget $\mathcal{B}$ after T AL cycles. At the end we will have a subset of videos $\mathcal{V}_s^T$ that has $\mathcal{F}_s^T = (\mathcal{F}_L^T, \mathcal{F}_U^T)$ frames, where $\mathcal{F}_L^T$ are annotated and $\mathcal{F}_U^T$ are unannotated frames. The proposed approach enables the

use of partial spatio-temporal annotation, utilizing both $\mathcal{F}_L^T$ and $\mathcal{F}_U^T$ for model training. We begin with an initial set of videos $\mathcal{V}_s^0 \subset \mathcal{V}$ with $\mathcal{F}_s^0 = (\mathcal{F}_L^0, \mathcal{F}_U^0)$ frames where $\mathcal{A}\%$ ($\mathcal{F}_L^0$) of these are annotated. We train the action detection model $\mathcal{M}^0$ using $(\mathcal{V}_s^0, \mathcal{F}_s^0)$ and use this trained model to select additional videos and frames using proposed AL to obtain new annotations. The proposed AL approach selects a diverse set of informative videos for annotation from $(\mathcal{V} - \mathcal{V}_s^0)$ which is added to $\mathcal{V}_s^0$ to obtain $\mathcal{V}_s^1$. Next we select $\mathcal{A}\%$ informative frames from the selected videos $\mathcal{V}_s^1$ for annotation. This iterative process is repeated till desired performance is met or the total budget is exhausted. An overview of the proposed approach is shown in Figure 2.

Video action detection: Video action detection requires spatial localization of the activity in each frame with temporal consistency of the predicted action location throughout the video. Most of the existing methods involve complex multi-stage training with dense frame-level annotations [21, 27, 66], making iterative training challenging due to large resource requirement and dependency on good region proposals [15, 22]. In this work, we utilize a simple one-stage approach which has state-of-the-art performance on action detection task and can be efficiently trained end-to-end using a single GPU [11]. We rely on the optimized version proposed in [31] to further reduce model complexity and the model is trained using margin-loss for classification and binary-cross entropy loss for action localization.

3.1. Hybrid active learning

The proposed hybrid active learning approach enables selection across unlabeled videos to identify diverse and important samples while also selecting limited frames within those samples for annotation; significantly reducing overall annotation cost. As shown in Figure 3, traditional sample selection approach simply selects and annotates entire sample, while intra-sample selection approach obtains frame-level annotations for all video samples. Sample selection does not take into account redundancy within a sample and intra-sample strategy on the other hand does not consider utility across samples and selects redundant samples, causing ineffective use of the annotation budget. We propose a hybrid approach that considers both intra-sample redundancy and inter-sample redundancy to select high utility frames and video samples. In addition, the proposed hybrid approach also integrates deep clustering to enable diversity along with informativeness while sample selection.

Inter-sample selection: In active learning, several approaches have been studied to estimate informativeness of a sample [13, 19, 45]. Motivated by the recent success of uncertainty-based approaches [14, 68], we focus on model uncertainty [14] to predict the utility of a video sample. In

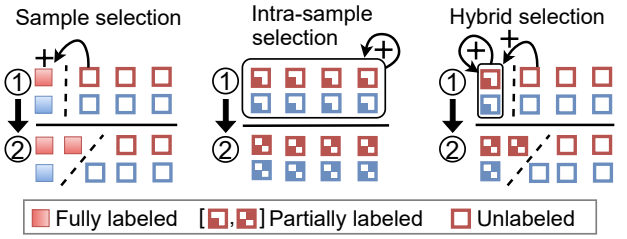

Figure 3. Overview of different active learning strategies for sample selection. We show a toy example for selection strategy as we add more annotatons to set 1 to obtain set 2. Sample selection approach takes unlabeled sample and annotates all frames in it. Intra-sample selects frames from all samples to annotate for the next set. Hybrid selects important samples and high utility frames to annotate for next set, significantly reducing overall annotation cost.

case of classification task, video-level classification uncertainty can be sufficient, however, video action detection also requires localization of actions on every frame of a video. Therefore, spatio-temporal localization also plays a crucial role in estimating samples utility. To take this into account, we rely on spatio-temporal uncertainty in our approach.

We consider uncertainty in model's prediction at pixel-level to compute spatio-temporal uncertainty. We rely on MC-dropout to compute model uncertainty [14, 26, 51] as it is more efficient in comparison with other approaches [1, 30, 44, 53]. The activity and non-activity region in a video will vary across action classes as well as across video samples. Therefore, uncertainty score based on all pixels in a video for sample utility will not be comparable across all unlabeled videos $\mathcal{V}_U$ for learning action detection. It will provide low uncertainty score for videos with short uncertain actions and long easy non-action regions which is not favorable for such videos. To overcome this issue, we propose to select limited frames in each video where we rank the video frames based on uncertainty and select the top $\mathcal{A}_t$ frames with high uncertainty. Given a pixel-level uncertainty $\mathcal{U}$, we compute the spatio-temporal uncertainty at video-level as,

$$V_{score} = \frac{1}{\mathcal{A}_t} \sum_{i=1}^{\mathcal{A}_t} \sum_{p=1}^{P} \mathcal{U}_{i,p} \tag{1}$$

where, $\mathcal{A}_t$ is the number of frames to select from each video in an AL iteration and P is the total pixels in each frame. The pixel-level uncertainty $\mathcal{U}$ is computed as,

$$\mathcal{U} = \frac{1}{R} \sum_{r=1}^{R} -log(\mathcal{M}(p, r)) \tag{2}$$

where, $\mathcal{M}(p)$ is the model's prediction of pixel p for each frame, averaged over R different runs. Uncertainty values for

$\mathcal{M}(p)$ below certain threshold (definite background) is set to 0. In our preliminary experiments, we observed that sample level classification uncertainty does not provide significant improvement over spatio-temporal uncertainty for sample utility. Therefore, we only utilize spatio-temporal uncertainty in our approach to determine sample informativeness for action detection.

Intra-sample selection: The informative videos selected in inter-sample selection $\mathcal{V}_{s-prime}^{t}$ are added to the existing set $\mathcal{V}_{s}^{t-1}$ to obtain $\mathcal{V}_{s}^{t}$. In intra-sample selection, we select frames with high utility from these videos $\mathcal{V}_{s-prime}^{t}$ for frame-level annotation. We rely on frame-level model uncertainty $U_f = \sum_i^I(\mathcal{U}_i)$ for all I pixels in a frame to estimate frame utility for action detection. Here $\mathcal{U}$ is pixel-level uncertainty as described in Equation 2. Since pixel-level uncertainty $\mathcal{U}$ is already computed for spatio-temporal uncertainty, intra-sample selection has no computation overhead.

Diverse sample selection: Model uncertainty can be used for sample selection focusing on their informativeness. However, it does not ensure diversity among selected videos and there can be redundancy in such a selection strategy. A simple solution to address this issue can be developed with the help of class labels. However, this will require additional annotations which defeats the purpose of saving annotation cost. We propose an implicit clustering approach which utilize latent video features and does not require additional annotations. More specifically, we use deep clustering [7] which learns the cluster representation for each category from the known labeled subset $\mathcal{V}_{s}^{0}$ and adapts the clusters as the latent features of each video changes during training.

To enable diverse sample selection, we model the relation between diversity of each unlabeled sample $\mathcal{V}_U$ with already labeled samples $\mathcal{V}_L$. The proposed clustering approach allows the model $\mathcal{M}$ to learn latent features $\mathcal{LF}$ which represent each sample in a cluster. The objective of the model $\mathcal{M}$ is to improve the latent features such that it is close to the corresponding cluster center for that sample. The clustering objective is defined as,

$$\min_{\theta} \mathcal{L}^{Cluster} = \sum_{i=1}^{N} \frac{\lambda}{2} ||\mathcal{LF}(x_i|M_\theta) - C_K(x_i)||^2 \qquad (3)$$

where, λ is a scaling term for the loss, θ is the parameters for model $\mathcal{M}$, $\mathcal{LF}$ is latent feature for sample x_i where $i \in [1, N]$ and C_K is the cluster center for sample x_i.

We first compute informativeness scores for each video in $\mathcal{V}_U$ using Equation 1, and then find cluster in $\mathcal{C} = [c_1, c_2, ...c_k]$ with K total clusters corresponding to each unlabeled video. The total number of videos to be selected in a cycle is constraint by current budget $\mathcal{B}_v$. We limit the samples selected per cluster such that the selection is proportional to the cluster size. For any cluster with n_c videos, we assign a budget of $n_c \times \mathcal{B}_v/N_U$, where N_U represents total number of unlabeled videos. The selection algorithm is further detailed in supplementary. We argue that nearby frames in a video will have similar model uncertainty and redundant utility. So, we avoid selecting nearby frames in intra-sample selection to ensure diversity while frame selection.

3.2. Training objective for partial label learning

Traditional video action detection method relies on actor annotation for each frame in order to train a model for action localization and classification [11, 28, 66]. However, in case of partial annotations it is not possible to train localization without annotations, which limits the use of these approaches directly. We propose a novel loss formulation which can effectively utilize partial annotations for localization.

Spatio-Temporal Weighted (STeW) loss: The partial spatio-temporal annotations can be converted into dense pseudo-labels with the help of interpolation [64]. However, these pseudo-labels can have errors due to motion of actor/camera in a video and temporal gap between the partial labels. We propose to use temporal continuity of actions to mitigate this issue and enable effective utilization of partial annotations. We hypothesize that actions have some temporal continuity across time which may vary with different actions. By leveraging this temporal continuity in a video, we compute spatio-temporal weight for each pixel independently which captures the confidence of a pseudo-label.

First, we compute the psuedo-labels using interpolation between the annotated frames as [64]. Next, we apply a spatio-temporal weight to suppress incorrect pseudo-labels. We compute the overlap of annotation for nearby frames and assign each pixel a weight based on the overall consistency which is given as,

$$\phi_f^{i,j} = Dist(f_a - f)\frac{1}{(W+1)} \sum_{w=f-W}^{f+W} f_w^{i,j}, \qquad (4)$$

where, weight ϕ of frame f with $i \times j$ pixels is combination of distance of frame f from nearest annotated frame f_a and average value of pixel i, j of nearby W frames. Our hypothesis is that the background and foreground should be consistent for most of the frame, except for the moving actions. The average value of nearby W pixels will give consistency value for each pixel, where we assign a weight of 1 for consistent background/foreground ($\leq P_{low}$ or $\geq P_{high}$) and average value for other inconsistent pixels. The final localization loss with spatio-temporal weight is computed as,

$$\mathcal{L}_l^{STeW} = \frac{1}{F} \sum_{f=1}^{F} \phi_f L_l^f, \qquad (5)$$

where, for a video with F frames, L_l^f is the BCE localization

loss for f^{th} frame and $\phi_f \in [0,1]$ is the pixel-wise spatio-temporal weighted mask from Equation 4 for f^{th} frame.

Overall training objective Our overall training objective is given as,

$$\min_{\theta} \mathcal{L} = \mathcal{L}^{Cluster} + \mathcal{L}_l^{STeW} + \mathcal{L}^{Cls} \quad (6)$$

where, θ is the model parameters, $\mathcal{L}^{Cluster}$ is cluster loss from Equation 3, $\mathcal{L}_l^{STeW}$ is detection loss from Equation 5 and $\mathcal{L}^{Cls}$ is the Margin-Loss for classification from [11].

3.3. Sampling budget and cost incorporation

For each stage of annotation we assume a fixed budget $\mathcal{B}$ which will be separated to annotate video label and to annotate frames in that video, given as $\mathcal{B}_v$ and $\mathcal{B}_f$ respectively. Annotating each video label will require a cost C_v since the annotator has to watch and identify the class. Similarly, annotating each frame with bounding-box or pixel-wise labels will require a cost C_f. Thus, for each stage we can only annotate videos and frames so that $C_v^{total} \leq \mathcal{B}_v$ and $C_f^{total} \leq \mathcal{B}_f$.

4. Experiments

Datasets: We evaluate our approach on **UCF-101-24** [59] and **J-HMDB-21** [25] action detection datasets. UCF-101-24 consists of 24 different action categories with spatio-temporal bounding-box annotations for 3207 untrimmed videos. J-HMDB-21 dataset has 21 categories with pixel-level spatio-temporal annotations for 928 trimmed videos. AVA [17] dataset has annotation on keyframes every second which makes it sparse and unsuitable to measure the effectiveness of the proposed approach (details in supplementary).

Evaluation metrics: We measure the standard frame-mAP and video-mAP scores for different thresholds to evaluate our model's action detection results following prior works [42]. The frame-mAP reflects the average precision of detection at the frame level for each class, which is then averaged to obtain the f-mAP [16]. The video-mAP reflects the average precision at the video level, which is averaged to obtain the v-mAP score

4.1. Implementation details

Active learning: We initialize our training with a set of videos $\mathcal{V}_L^0$ with class label and $\mathcal{A}\%$ annotated frames within those videos selected at random. We use **K=5** centers for clustering (analysis on varying K in supplementary) and **R=10** forward passes per video. For each stage, we select $v\%$ videos for annotation based on budget $\mathcal{B}_v, \mathcal{B}_f$, where the videos are given class label and $\mathcal{A}\%$ of their frames are annotated and added to $\mathcal{V}_L^0$. We repeat this until total budget is exhausted or desired performance is achieved.

		UCF-101-24				J-HMDB-21	
$\mathcal{A}\%$	$V\%$	**v-mAP**	**f-mAP**	$\mathcal{A}\%$	$V\%$	**v-mAP**	**f-mAP**
0.25	5	45.0	50.1	0.15	5	4.7	27.2
0.50	10	54.3	55.6	0.30	10	41.6	45.3
0.75	15	57.6	59.4	0.45	15	52.5	54.8
1.00	20	61.8	61.6	0.60	20	56.0	60.5
1.25	25	65.5	65.6	0.75	25	57.6	60.9
1.50	30	67.2	66.9	0.90	30	58.3	61.7
2.00	40	68.6	68.5	1.20	40	61.3	62.7
2.50	50	69.2	69.3	1.50	50	63.7	64.0
5.00	80	72.2	72.1	5.40	80	71.5	72.8
90	90	73.6	73.0	90	90	73.1	73.0
100	100	75.2	74.0	100	100	75.8	74.9

Table 1. Evaluation of the proposed method on **UCF-101-24** and **J-HMDB-21** for [v-mAP, f-mAP] @ 0.5 IoU. We increase the amount of samples and frames in each stage using the proposed approach and compare with fully-supervised approach. $\mathcal{A}\%$ is percent of annotated frames.

Training details: All our experiments are performed using PyTorch [41] on a single Nvidia Quadro 5000 GPU. The scores are average of 3 different runs. We adapt the video action detection model from [11] and use 2D capsules and I3D encoder [8] following [31], with pretrained weights from the Charades dataset [56]. The network is trained using Adam optimizer [29] with learning rate $5e-4$ and batch size 8. $P_{low} = 0.1$ and $P_{high} = 0.9$ is set empirically. We use random crop and horizontal flip for video augmentation during training. Interpolation is done using linear point interpolation for bounding-box (UCF-101-24) and CyclicGen [37] for pixel-wise (JHMDB). We compute uncertainty based on dropout during inference following [14]. We don't perform any hyperparameter tuning and use same set of parameter settings for all our experiments on both the datasets.

4.2. Baseline methods

We compare the proposed approach with several baselines to demonstrate its effectiveness. We develop two non-parametric selection method using random and equidistant frame selection (both using random video selection). We also use prior AL methods for object detection in images as baselines. We use uncertainty-based AL [14] and entropy-based AL [1] for scoring each frame and do sample selection. All baselines use same action detection backbone as ours.

4.3. Results

We show that our iterative AL approach is able to improve results in each step and use only a fraction of the annotations to perform close to fully-supervised approach with 90% annotations *(v-mAP@0.5: 72.2 vs 73.6 (UCF-101-24), 71.5 vs 73.1 (J-HMDB-21))* in Table 1. We also perform detailed comparisons with 4 baselines in Table 2. We further com-

Method	$\mathcal{A}\%$	UCF-101-24		J-HMDB-21	
		v-mAP	f-mAP	v-mAP	f-mAP
Random	1%	52.6	54.1	36.6	42.1
Equi.	1%	53.3	55	38.1	43.5
Entropy [1] †	1%	52.2	53.5	40.7	49.0
Uncertainty [14] †	1%	44.0	46.7	46.0	47.9
Our	1%	61.8	61.6	58.6	61.9
Random	5%	67.5	67.3	69.3	70.1
Equi.	5%	67.2	67.0	70.0	70.4
Entropy [1] †	5%	71.3	70.2	70.7	70.8
Uncertainty [14] †	5%	69.7	68.2	69.0	69.3
Our	5%	72.2	72.1	71.3	72.7

Table 2. Comparison of the proposed approach with various baseline methods. All baseline methods use same action detection backbone as ours. † is modified for video action detection using public code. $\mathcal{A}\%$ is total annotations.

pare our approach with previous weakly-supervised action detection approaches on both the datasets in Table 3 and 4.

Comparison with baselines: Table 2 shows comparison of our method with random, equidistant, entropy-based [1] and uncertainty-based [14] AL baselines for UCF-101-24 and J-HMDB-21. We report the f-mAP and v-mAP scores at 1% and 5% total annotations. Random and equidistant give an idea of non-parametric sample selection where the videos are selected at random and the frames are selected at random or equidistant. We notice that these baselines give lowest scores. Then we compare with other AL baselines using [1, 14]. Since these are image-based, they are not well suited for frame ranking in videos as reflected by their scores. [1] ignores nearest 5 frames for each selection, but this still does not work as well as proposed diverse selection. Since these prior AL baselines don't have notion of similarity/distance for videos, we see that random performs comparably. In contrast, our approach gives best performance, highlighting the impact of cluster based diverse sample selection.

Comparison with weakly supervised approach: Our cluster based video and frame selection approach selects limited samples and can also be compared with prior weakly supervised methods for video action detection. Prior weakly supervised methods rely on multiple instance learning [3, 39, 40] or instance learning [64], paired with off-the-shelf actor detector or user-generated points to create GT annotations for training. These rely on multiple external components or require user to annotate points in each frame, reducing their practical use. Some methods are less involved with built-in detector branch [12] but suffer from noisy annotations. [9] applies discriminative cluster approach to match generated actor tubes with video label with partially annotated frames. [67] combines multiple actor detectors to build

Method	$\mathcal{A}\%$	f-mAP@	v-mAP@			
		0.5	0.1	0.2	0.3	0.5
Mettes et al. [40]	V	-	-	37.4	-	-
Escorcia et al. [12]	V	-	-	45.5	-	-
Zhang et al. [67]	V	30.4	62.1	45.5	-	17.3
Arnab et al. [3]	V	-	-	61.7	-	35.0
Mettes et al. [39]	P	-	-	41.8	-	-
Cheron et al. [9]	P	-	-	70.6	-	38.6
Weinz. et al. [64]	1.1%	-	-	57.1	-	46.3
Weinz. et al. [64]	2.8%	63.8	-	57.3	-	46.9
MixMatch [5]	S-20%	20.2	-	60.2	-	13.8
Pseudo-label [32]	S-20%	64.9	-	93.0	-	65.6
Co-SSD(CC) [24]	S-20%	65.3	-	93.7	-	67.5
Kumar et al. [31]	S-20%	69.9	-	95.7	-	72.1
Ours	1%	61.6	98.1	95.9	88.9	61.8
Ours	5%	72.1	98.1	96.1	91.2	72.2

Table 3. Comparison with state-of-the-art weakly-supervised methods on UCF-101-24. We evaluate our approach on v-mAP and f-mAP scores using only 1% and 5% total frame annotations. 'V' uses video-level annotations and 'P' uses a fraction of the mixed annotation. 'S' denotes SSL methods. We report [64] with their scores for 2 (1.1%) and 5 (2.8%) frames annotated per video.

Method	$\mathcal{A}\%$	f-mAP@	v-mAP@			
		0.5	0.1	0.2	0.3	0.5
Zhang et al. [67]	V	65.9	81.5	77.3	-	50.8
Weinz. et al. [64]	6%	50.7	-	-	-	58.5
Weinz. et al. [64]	15%	56.5	-	-	-	64.0
MixMatch [5]	S-30%	7.5	-	46.2	-	5.8
Pseudo-label [32]	S-30%	57.4	-	90.1	-	57.4
Co-SSD(CC) [24]	S-30%	60.7	-	94.3	-	58.5
Kumar et al. [31]	S-30%	64.4	-	95.4	-	63.5
Ours	1%	61.9	99.0	96.8	91.5	58.6
Ours	5%	72.7	99.1	97.3	94.8	71.3

Table 4. Comparison with state-of-the-art semi-supervised methods on J-HMDB-21 using only 1% and 5% total frames annotation. 'V' uses video-level class annotations. 'S' denotes SSL method. We report [64] with their scores for 2 (6%) and 5 (15%) frames annotated per video.

stronger GT annotations, relying heavily on external components. Our approach doesn't rely on external detection components and uses simple iterative approach to select useful limited samples. This allows our method to be easily used for training. Table 3 and 4 shows comparative scores with prior weakly-supervised methods.

4.4. Ablations

Effect of clustering: We evaluate the effect of clustering for video selection in our approach in Figure 5. The selection approach without clustering simply selects *top-k* videos for further annotation, which ends up selecting some similar samples as it does not take diversity into account. Clustering increases sample diversity, as seen in Figure 4, which

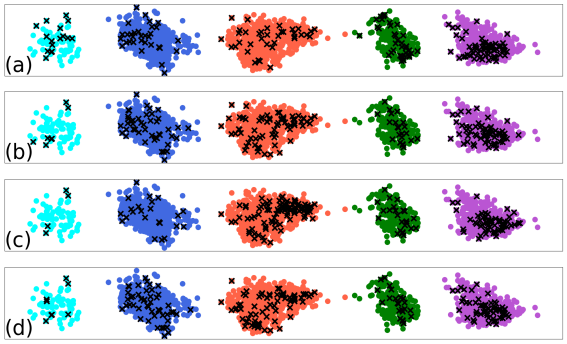

Figure 4. Visual representation of samples selected using (a) proposed *Clustering-Aware Uncertainty Scoring (CLAUS)*, (b) entropy, (c) uncertainty and (d) random selection methods using **x** marks. We get latent features of the videos from same iteration using same model and project them after PCA reduction. The clusters are from our clustering method and only for visual demonstration in (b), (c) and (d). We observe that our approach has diverse and even sample selection from different clusters while (b), (c) and (d) often selects samples closer to each other in terms of representation.

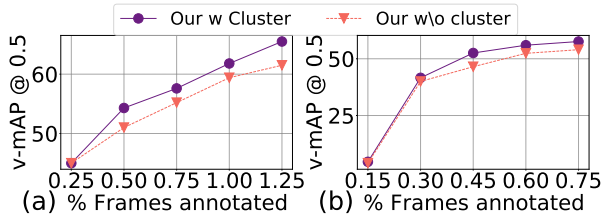

Figure 5. Comparison of our approach with and without clustering based selection for UCF-101-24(a) and J-HMDB-21(b).

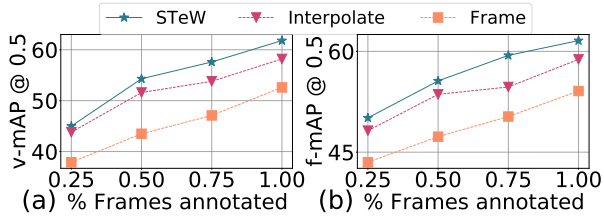

Figure 6. Comparison of proposed *STeW* loss with different loss variations combined with our ***CLAUS*** selection to train the video action detection network for UCF-101-24 dataset.

improves overall performance compared to non-clustering selection for both datasets as shown in Figure 5.

Effectiveness of STeW loss: To evaluate the effect of our proposed *STeW* loss, we train the action detection network using simple *frame* loss and *interpolation* loss for UCF-101-24 dataset. *Frame* loss only computes loss for the annotated

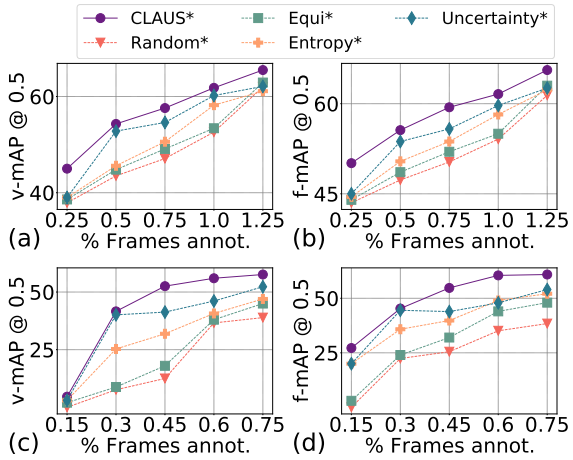

Figure 7. Evaluating various scoring methods for AL based annotation increments. * uses **our *STeW Loss*** for all selection approaches on UCF-101-24(a-b) and J-HMDB-21(c-d).

frame and ignores the pseudo-labels while *interpolation* loss simply computes loss for all real and pseudo-labels equally. We use the same AL algorithm for all the approaches and show the result for UCF-101-24 for different steps in Figure 6. With less than 1% frames annotated, we see that *Frame* loss is not able to learn detection as well as *interpolation* and *STeW* loss. With the pseudo-labels created by interpolating the annotated frames, we see an increase in performance across all steps with both *interpolation* and *STeW* loss. Furthermore, the proposed *STeW* loss gives more importance to real frames and reduces the impact of the pseudo-labels that are inconsistent, performing best among all loss variations.

Effectiveness of CLAUS scoring: We also evaluate different scoring functions (random, equidistant, entropy [1], uncertainty [14]) paired with the proposed *STeW* loss in Figure 7. Proposed *CLAUS* method is the only one that selects diverse samples based on global utility and is able to perform best compared to other scoring functions.

4.5. Discussion and analysis

Cost analysis: Figure 8(a-b) compares cost to performance relation of our method and random selection. While having more annotation generally improves performance, our method selects diverse and important frames compared to random selection, resulting in significantly improved model in each step for the same cost. We further take the final model and evaluate per class performance for our and random selection in Figure 9. We outperform random selection for most classes while having fewer samples selected and give priority to select more samples for certain harder classes.

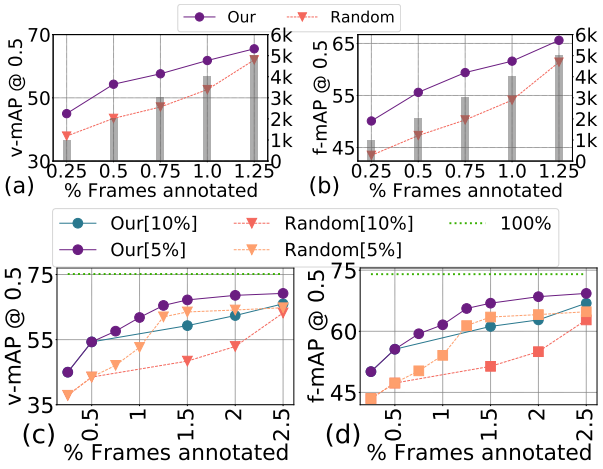

Figure 8. (a-b) Performance evaluation of our method with random selection baseline on UCF-101-24 for various sample annotation percent. The cost of annotation for each step is shown by the shaded bars, with the cost value in the right axis in thousands. (c-d) Performance difference for increasing sample and frame annotations [5%] vs increasing only frame annotations [10%] on UCF-101-24. Increasing both sample and frames at 5% increment adds diversity compared to only increasing frames, giving better scores.

Sample vs frame increment: We evaluate effect of increasing only samples with a constant frame annotation rate of 5% and increasing both samples and frames annotation. Our goal is to get maximum performance gain with lowest cost. Increasing only samples with constant frame annotation rate has lower annotation cost than increasing both samples and frames for the same step. We show the results in Figure 8(c-d); having more training variation by adding only samples is more cost effective and has better performance than having more frames annotated for the same samples with higher cost. Interestingly, even random sampling that increases sample diversity performs better than our sampling with more frames, showing that sample diversity is an important factor in the selection process.

Class vs clustering diversity: While samples from different classes add diversity, too many samples for easy classes will also add redundancy. Figure 9 shows that random approach has class balanced selection but performs below *CLAUS* as *CLAUS* reduces redundant samples from same class and prioritizes difficult and diverse samples.

Selection strategy analysis: We compare selection using proposed hybrid method against classical approaches for the same annotation budget. *Inter selection* assumes each video is fully annotated and randomly selects videos for given budget, thereby selecting fewer videos as more budget is used to annotate all frames. *Intra selection* assumes each video

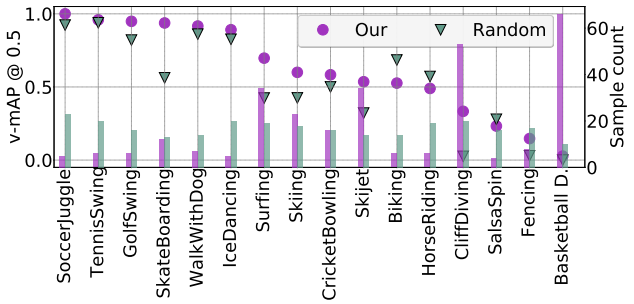

Figure 9. Analysis on performance across classes with varying amount of annotations. The scatter plot with markers on left axis shows v-mAP scores @ 0.5 IoU of our method against baseline random method on 16 action classes for UCF-101-24. The bar plot with markers on right axis shows per class sample distribution.

of the dataset is annotated for at least 1 frame, spreading the budget over all videos. We show this comparison in Figure 1; our proposed method consistently scores better with both hybrid selection and random selection. Inter selection simply exhausts the budget in redundant frames from fewer videos and performs worst. Intra selection does perform close to our-with-random baseline due to larger sample variation.

5. Conclusion

In this work we present a novel hybrid AL strategy for reducing annotation cost for video action detection. Our hybrid approach uses clustering-aware strategy to select informative and diverse samples to reduce sample redundancy while also doing intra-sample selection to reduce frame annotation redundancy. We also propose a novel *STeW loss* to help the model train with limited annotations, removing the need for dense annotations for video action detection. In contrast to traditional AL approach, our proposed hybrid approach adds more annotation diversity at the same cost. We evaluate the proposed approach on two different action detection datasets demonstrating its effectiveness in learning from limited labels with minimal trade-off on the performance.

6. Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (Intelligence Advanced Research Projects Activity) via 2022-21102100001 and in part by University of Central Florida seed funding. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the US Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López. Active learning for deep detection neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3672–3680, 2019. 2, 3, 5, 6, 7

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2

[3] Anurag Arnab, Chen Sun, Arsha Nagrani, and Cordelia Schmid. Uncertainty-aware weakly supervised action detection from untrimmed videos. In *European Conference on Computer Vision*, pages 751–768. Springer, 2020. 1, 2, 6

[4] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019. 2

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 6

[6] Zalán Bodó, Zsolt Minier, and Lehel Csató. Active learning with clustering. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pages 127–139. JMLR Workshop and Conference Proceedings, 2011. 2

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 4

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 2, 5

[9] Guilhem Chéron, Jean-Baptiste Alayrac, Ivan Laptev, and Cordelia Schmid. A flexible model for training action localization with varying levels of supervision. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 950–961, 2018. 1, 2, 6

[10] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022. 1

[11] Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Videocapsulenet: A simplified network for action detection. In *Advances in Neural Information Processing Systems*, pages 7610–7619, 2018. 2, 3, 4, 5

[12] Victor Escorcia, Cuong D Dao, Mihir Jain, Bernard Ghanem, and Cees Snoek. Guess where? actor-supervision for spatiotemporal action localization. *Computer Vision and Image Understanding*, 192:102886, 2020. 1, 2, 6

[13] Alireza Fathi, Maria Florina Balcan, Xiaofeng Ren, and James M Rehg. Combining self training and active learning for video segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2011. 2, 3

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2, 3, 5, 6, 7

[15] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019. 2, 3

[16] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 759–768, 2015. 5

[17] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 5

[18] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 1, 2

[19] Fabian Caba Heilbron, Joon-Young Lee, Hailin Jin, and Bernard Ghanem. What do i annotate next? an empirical study of active learning for action localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 199–216, 2018. 1, 2, 3

[20] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008. 2

[21] Rui Hou, Chen Chen, and Mubarak Shah. An end-to-end 3d convolutional neural network for action detection and segmentation in videos. *arXiv preprint arXiv:1712.01111*, 2017. 3

[22] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *IEEE International Conference on Computer Vision*, 2017. 1, 2, 3

[23] Alexandros Iosifidis, Anastasios Tefas, and Ioannis Pitas. Semi-supervised classification of human actions based on neural networks. In *2014 22nd International Conference on Pattern Recognition*, pages 1336–1341. IEEE, 2014. 1

[24] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32, 2019. 6

[25] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 5

[26] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379. IEEE, 2009. 2, 3

[27] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International*

Conference on Computer Vision, pages 4405–4413, 2017. 1, 2, 3

[28] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Joint learning of object and action detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4163–4172, 2017. 4

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[30] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019. 2, 3

[31] Akash Kumar and Yogesh Singh Rawat. End-to-end semi-supervised learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14700–14710, 2022. 1, 3, 5, 6

[32] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 6

[33] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 303–318, 2018. 2

[34] Xin Li and Yuhong Guo. Adaptive active learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866, 2013. 1, 2

[35] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *European Conference on Computer Vision*, pages 68–84. Springer, 2020. 1

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[37] Yu-Lun Liu, Yi-Tung Liao, Yen-Yu Lin, and Yung-Yu Chuang. Deep video frame interpolation using cyclic frame generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8794–8802, 2019. 5

[38] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6122–6131, 2019. 2

[39] Pascal Mettes and Cees GM Snoek. Pointly-supervised action localization. *International Journal of Computer Vision*, 127(3):263–281, 2019. 1, 2, 6

[40] Pascal Mettes, Cees GM Snoek, and Shih-Fu Chang. Localizing actions from video labels and pseudo-annotations. *arXiv preprint arXiv:1707.09143*, 2017. 1, 2, 6

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[42] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In *European conference on computer vision*, pages 744–759. Springer, 2016. 2, 5

[43] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *Proceedings of the ieee international conference on computer vision*, pages 5179–5188, 2017. 2

[44] Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[45] Ameya Prabhu, Charles Dognin, and Maneesh Singh. Sampling bias in deep active classification: An empirical study. *arXiv preprint arXiv:1909.09389*, 2019. 2, 3

[46] Aayush J. Rana and Yogesh S. Rawat. We don't need thousand proposals: Single shot actor-action detection in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2960–2969, January 2021. 2

[47] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[48] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021. 2

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2

[50] Mamshad Nayeem Rizve, Ugur Demir, Praveen Tirupattur, Aayush Jung Rana, Kevin Duarte, Ishan R Dave, Yogesh S Rawat, and Mubarak Shah. Gabriella: An online system for real-time activity detection in untrimmed security videos. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4237–4244. IEEE, 2021. 1

[51] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2021. 3

[52] Mamshad Nayeem Rizve, Navid Kardan, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 382–401. Springer, 2022. 2

[53] Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. Towards realistic semi-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 437–455. Springer, 2022. 3

[54] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 2

[55] Burr Settles. Active learning literature survey. 2009. 2

[56] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 5

[57] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10399, 2021. 1

[58] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 2

[59] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 5

[60] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4489–4497. IEEE, 2015. 1

[61] Carl Vondrick and Deva Ramanan. Video annotation and tracking with active learning. *Advances in Neural Information Processing Systems*, 24:28–36, 2011. 1

[62] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016. 2

[63] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3164–3172, 2015. 2

[64] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*, 2016. 1, 4, 6

[65] Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 917–926, 2009. 2

[66] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2019. 1, 2, 3, 4

[67] Shiwei Zhang, Lin Song, Changxin Gao, and Nong Sang. Glnet: Global local network for weakly supervised action localization. *IEEE Transactions on Multimedia*, 22(10):2610–2622, 2019. 2, 6

[68] Guang Zhao, Edward Dougherty, Byung-Jun Yoon, Francis Alexander, and Xiaoning Qian. Uncertainty-aware active learning for optimal bayesian classifier. In *International Conference on Learning Representations (ICLR 2021)*, 2021. 2, 3

[69] Javad Zolfaghari Bengar, Abel Gonzalez-Garcia, Gabriel Villalonga, Bogdan Raducanu, Hamed Habibi Aghdam, Mikhail Mozerov, Antonio M López, and Joost van de Weijer. Temporal coherence for active learning in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2